

PATENT APPLICATION

Computational Protein Probing to Identify Binding Sites

Frank Guarnieri
1742 West Eleventh Street
Brooklyn, NY 11223-1142

Computational Protein Probing to Identify Binding Sites

The present application claims the priority of US Provisional Application No.
5 60/101,521, filed September 23, 1998.

The present invention relates to methods of identifying binding sites on proteins, methods for identifying classes of compounds suitable for binding a protein, and methods of conducting experiments to identify compounds that interact with a protein to affect a biological process.

10 Determinations of protein structures have to date been conducted by isolating crystals of the protein of interest, and analyzing structure by X-ray crystallography. Typically, the protein has been co-crystallized with heavy metal component, or subjected to multiple co-crystallizations, with the heavy metal providing a reference for solving the crystallographic data.

15 With a determination of the structure of a protein, or the structure of another macromolecule having significant tertiary structure, such as a DNA or RNA, workers often seek to identify the binding sites that are or may be of significance to a biological process, such as an enzyme active site or a site for interacting with another macromolecule or with itself. Computational efforts have been focused on efforts to
20 sample the surface of a molecule to find good fits with known binding agents. These methods have had modest success, and are dependent on knowledge of (a) the structure of good binding agents and, often, (b) the function of the protein. A more traditional approach has sought to co-crystallize binding substances with the macromolecule to identify binding sites. With the binding site identified, educated guesses can be made as
25 to new molecules that could bind the site. These educated guesses can guide synthetic methods, including combinatorial chemistry methods, to make and test new molecules. When such prospective binding agents prove effective binding agents, and possibly are also found effective in an appropriate biological model, the structural correlations drawn from the results can be tied to information about the binding site to make still further
30 inferences about the structure important to a biological function. This co-crystallization approach depends on an initial knowledge of active agents, and is experimentally difficult and time consuming.

The present inventor has found a method of identifying, from a three-dimensional structural solution of a macromolecule, the binding sites for molecules. The structural solution used as the basis for the method can be derived from crystallography, spectroscopic analyses such as NMR, computational derivations, or any other method of determining the structure of a macromolecule. The method does not require or typically use information on the function of the macromolecule, as the method avoids subjective biases and instead depends purely on physical parameters. Further, the method can be refined further to narrow the possible choices of binding sites and identify the functionalities, i.e., organic fragments or "ORFs," that effectively interact with the binding site(s). The data obtained for ORFs further identifies the orientations of the functionalities useful in a candidate binding agent, thereby providing a tool for searching chemical databases to identify candidate binding agents. Where the methods described herein identify more than one potential binding site, the data generated through these methods can be used to energetically rank the binding sites, and thereby quantitatively determine which site has the potential to more strongly bind molecules.

The computational method described here generates maps of binding site preferences that are nearly identical with maps produced by compiling data generated by traditional methods, but with one important difference – the experimentally produced data took many years to produce while the data produced as described herein can be produced in no more than a few weeks. The invention provides an important development in unbiased simulation methods for predicting the character of agents that bind to biological macromolecules to affect the function of the macromolecules.

Summary of the Invention

In one embodiment, provided is a method of identifying binding sites on a macromolecule comprising: (a) for at least one organic fragment (ORF), conducting, at separate values of parameter *B*, two or more simulated annealing of chemical potential calculations using the ORF as the inserted solvent; and (b) comparing converged solutions from step (a) to identify first locations at which the relevant ORF is strongly bound, thereby identifying candidate sites for binding ligand molecules. In one preferred aspect, the method further comprises: (c) identifying clusters of sites that strongly bind an ORF. In another preferred aspect, the method further comprises: (d) conducting steps

(a) and (b) for each of two or more ORFs and identifying clusters where two or more distinct ORFs bind. Preferably, a cluster that binds three or more distinct ORFs is identified. The method can identify further functionalities that contribute to the binding of bioactive agents by reducing the binding stringency in the vicinity of a cluster to
5 further identify elements that would contribute to the binding of a bioactive agent.

In another preferred aspect, the method further comprises: (e) conducting, at separate values a measure of chemical potential, two or more simulated annealing of chemical potential calculations using water as the inserted solvent; (f) comparing converged solutions from step (c) to identify locations at which water is strongly bound,
10 thereby identifying locations on the protein which are not candidate sites for binding ligand molecules; and (g) identifying first locations that are not water locations.

In still another preferred aspect, the simulated annealing of chemical potential calculations comprise multiple steps of sampling, and wherein in a number of steps of the sampling the ORFs position is changed by a small amount and the resulting new position
15 is accepted or rejected based on the change in energy as a result of the change attempted.

Further provided is a method of identifying the chemical characteristics of compounds that bind a macromolecule comprising examining the functionalities and relative orientations of the ORFs found in a cluster pursuant to the binding site identifying method outlined above.

Also provided is a method of conducting combinatorial chemistry to identify
20 compounds that interact with a macromolecule comprising: (a) identifying classes of reactants that are modeled by the functionalities of the ORFs found in a cluster pursuant to the binding site identifying method of macromolecule; (b) designing a combinatorial synthetic protocol that calls for two or more synthetic procedures that react reagents of at
25 least two of the classes identified in step (a); and (c) conducting the combinatorial synthetic protocol to create candidate binding molecules.

Further provided is a method of conducting a bioactive agent discovery process comprising: (a) from a group of established combinatorial synthetic protocols or a collections of chemical compounds or pools of chemical compounds, identifying those
30 members of the group that provide a high density of compounds that meet for a macromolecule selection criteria identified from the binding site identifying method of macromolecule; and (b) conducting binding or functional assays to identify compounds

obtained from the identified collections or protocols which bind or affect the function of the macromolecule.

BRIEF DESCRIPTION OF THE DRAWINGS

5 **Figure 1A** illustrates a solved crystal structure, while **Figure 1B** displays the structure with a grid is imposed.

Figures 2A-2D display the method of the invention applied to the crystallographic solution of elastase, the method can be exemplified using methanol as the ORF.

10 **Figures 3A and 3B** show the combined results for several ORFs bound to elastase after simulations at relatively low *B* values, with the results in **Figure 3B** filtered to identify clusters of these bound ORFs.

Figure 3C shows the two clusters of **Figure 3B** which remain after excluding strong water binding sites, and **Figure 3D** shows the one cluster that remains after
15 extending the analysis to another ORF; **Figure 3E** shows the analysis extended to still a further ORF.

 The panels of **Figure 3F** compare the simulation results to a co-crystallography result.

 Illustrated in **Figure 4A** are the amide binding sites extracted from the data of six
20 co-crystallization experiments with elastase and known ligands; and illustrated in **Figure 4B** is a cluster of the highest affinity amide binding sites determined by the simulation method of the invention.

 Illustrated in **Figure 4C** are the amide ORFs of **Figure 4B** plus amides which are in the vicinity of the cluster but which appear in the simulation at second highest affinity
25 binding values.

 In **Figures 5A and 5B**, solutions obtained with co-crystals of elastase inhibitors are compared with data obtained by the methods herein described.

Figures 6A and 6B show the surfaces of elastase involved in binding ligands as indicated by the crystallographic data, **Figure 6A**, and as indicated by the solutions
30 obtained the method described herein, **Figure 6B**.

09182267 103099

Figure 7 shows a schematic illustration of the type of titrations for water binding to a macromolecule that can be used to help identify a level of relatively strong water binding.

5 GLOSSARY

The following terms shall have, for the purposes of this application, the respective meaning set forth below.

- **"Bioactive agent"** refers to a substance such as a chemical that can act on a cell, virus, organ or organism, including but not limited to drugs (i.e. pharmaceuticals) to create a change in the functioning of the cell, virus, organ or organism. In a preferred embodiment of the invention, the method of identifying bioactive agents of the invention is applied to organic molecules having molecular weight of about 600 or less or to polymeric species such as peptides, proteins, nucleic acids, proteoglycans and the like. A bioactive agent can be a medicament, i.e. a substance used in therapy of an animal, preferably a human.
- **"Cluster of free grid points"** refers to free grid points that are within a "cluster" in that, relative to a given ORF, there is a sufficient number of nearby or adjacent free grid points to allow a reasonable probability that the ORF could be inserted at the cluster. Thus, the cluster of free grid points for H₂O must be defined to identify all volumes at the surface or interior of a macromolecule that could accommodate H₂O – though the selection criteria should err to identifying some volumes that do not accommodate H₂O, as needed to assure that all appropriate volumes are sampled in the simulation process. A cluster of free grid points is defined differently depending on the size of the ORFs (e.g., compare H₂O and benzene) and the spacing of the grid.
- A **"cluster of ORF binding sites"** typically refers to a pattern of closely located or superimposed sites that bind ORFs with sufficient affinity to merit further consideration.
- **"Collection of chemical compounds"** refers to any collection of compounds collected or organized with the intention that they can be examined to identify bioactive agents (e.g., having a biological activity measured directly or through a surrogate for biological activity such as binding to a macromolecule or interfering with a function of a macromolecule). The collection can be prepared from a collection of simpler molecules (which can be bound to a support) by a chemical scheme designed to generate a diversity

of chemicals. Collections of this latter type are often referred to as "combinatorial libraries."

- **"Free grid points"** refers to grid points (which are discussed below) which are, for a given accepted definition of atomic radius, "free" in that they do not fall within the atomic radii of the mapped atoms of the relevant macromolecule.
- **"Macromolecule"** refers to a molecule or collection of molecules which has a time-averaged tertiary structure. Thus, while the term typically refers to proteins, ribonucleic acids, structures formed of both nucleic acid and protein, carbohydrates, structures formed of two or more of the aforementioned, and the like, it can also refer to structures formed with other molecules including lipids. Macromolecules are used in the method described herein with reference to maps of their tertiary structure. Such maps are typically generated by X-ray diffraction studies, which have generated maps for thousands of macromolecules. However, maps can be produced by other methods such as computational methods or computational methods supplemented by other data such as NMR data. While computational methods have been difficult to apply, recent studies appear to have achieved some successes.
- **Organic fragments or "ORFs"** are molecules or molecular fragments that can be used to model one or more modes of interaction with a macromolecule, such as the interactions of carbonyls, hydroxyls, amides, hydrocarbons, and the like.
- **Water locations** are locations at which water is strongly bound, meaning, in one embodiment, for example locations where the simulation indicates water remains bound when the simulation is run at values of B that are equal to or less than the B value for the transition point indicating those water molecules that are strongly influenced by the macromolecule. Illustrated in **Figure 7** is a conceptualization of the titration of simulated bound water molecules with decreasing values of B , a parameter described further below. A transition point indicates water molecules that are strongly influenced by the macromolecule. A B value less than or equal to that at the transition point can be designated as defining water binding of sufficient strength to render competitive binding by another molecule unlikely, as illustrated by point **SB** in the illustration. Typically, for a water soluble protein, this point **SB** is selected so that about 100 to about 50 water molecules remain bound for a 50 kd protein.

DETAILED DESCRIPTION OF THE INVENTION

The simulation process of the present invention works by artificially inserting a given ORF at an unbiased sampling of all the sites on or within a macromolecule structure where such ORF can, as a practical matter, reside. These sites can be termed the "sampling sites." Typically, a schedule of simulations for each of a number of ORFs are run, with each simulation run at a separate value of a parameter B , which is related to the excess chemical potential. The schedule provides for simulations conducted at each of a number of B values, typically ranging from 10 to about -15. In each simulation at a given value of B , the simulation assesses at each step of the simulation whether the insertion of the ORF at a given site shall be accepted or rejected, with the assessment based on a grand canonical ensemble probability density function. At each step of the simulation, the algorithm models the insertion of the ORF at the site. A forced bias canonical probability density function is used to translate and rotate the ORF in small steps (e.g., $\pm 0.2\text{\AA}$, $\pm 30^\circ$) to identify an energy minimized insertion given the simulation parameters in place at the time of the simulation step. The probability of the insertion is then determined from the grand canonical ensemble probability density function, and the ORF can be represented as resident at the site by a random number generating protocol weighed to the probability value. Alternative methods for choosing to make this representation, such as applying cutoff values for when to represent the insertion or not, can also be applied, but are less favored. Typically, following a successful insertion, the subsequent deletion attempts at the site are with the previously identified translated and rotated ORF, and this translated and rotated ORF is used until a deletion attempt succeeds. The simulation is typically conducted for a large number of steps, such as 2×10^6 steps, with the majority of the steps, e.g., 1.5×10^6 required to "equilibrate" the simulation so that the number of accepted insertions is equal to the number of deletions on average.

By taking a large number of unbiased samplings at each sample site over the latter course of the simulation, such as after every 200 steps of iterations after equilibrium is achieved, an occupation probability of the ORF residing at that sample site at the given value of B can be assessed. The occupancy as an overall result of the method can then be determined based on this probability, for example with a random number protocol

making the representation based on its probability. The degree to which the ORF is translated or rotated can also be represented based on the probability of such translations and rotations.

For each ORF, simulations are run at each of a number of values of a measure of excess chemical potential, such as B . Thus, as this value lowers, the retention of an ORF at a given sampling site is an indication of high relative binding affinity.

The sampling sites are typically arrived at by creating a grid as illustrated in Figure 1. Figure 1 illustrates a solved crystal structure (Figure 1A) on which a grid is imposed (Figure 1B). For example, the grid can have about $\frac{1}{2}$ Å to about 1 Å spacing, with the grid intersection points defining the candidates for sampling sites. The spacing of the grid is preferably selected to be less than the smallest cross-section of the ORF. The spacing is typically selected to be small enough in relation to the size of the ORF so that the probability that free volumes that could define free grid point clusters have sufficient free grid points to allow useful sampling as described below. Such relatively small spacing minimizes the chance that the selection of how to orient the grid will bias the algorithm against identifying certain ORF binding preferences. The sampling sites are selected from sites that are unoccupied by the macromolecule (Figure 1B). A final elimination of "grid bias" is achieved by varying the test insertion points away from strict initial insertion at grid points, as described below.

The sampling sites are limited to those sites having enough adjacent volume free of the macromolecule to allow the ORF to be inserted. For example, the sampling sites can be limited to grid points within an open area of at least about $2\text{Å} \times 2\text{Å} \times 2\text{Å}$ ($= 8\text{Å}^3$ or 0.008nm^3) or about $2.5\text{Å} \times 2.5\text{Å} \times 2.5\text{Å}$ ($= 15.6\text{Å}^3$ or 0.0156nm^3) or, for water, about $2.2\text{Å} \times 2.2\text{Å} \times 2.8\text{Å}$. The grid points can be selected for those free grid points that are within a cluster of free grid points, such as, for example, a cluster of 3, 4, 5, 6, 7, 8 or more free grid points, depending on the size of the ORF and the spacings of the grid.

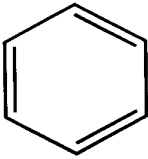
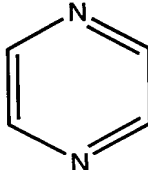
In one preferred embodiment, the ORF is not necessarily initially inserted exactly at the grid points, but instead at a random sampling of insertion points within a short distance of the grid points, such as points within a sphere shape centered at the grid point and having a diameter of about some percentage, such as 10%, of the grid spacing, or within a box shape centered at the grid point having width, length and height of about such a percentage of the grid distance. As discussed above, this "wobble" in the initial

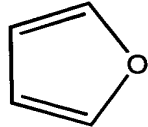
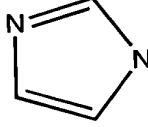
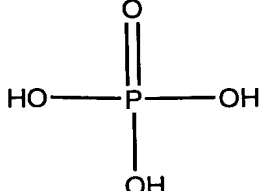
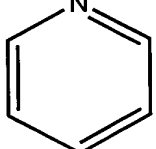
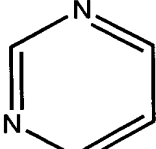
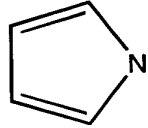
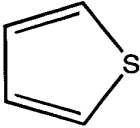
insertion point helps eliminate grid bias where the placement of the grid happens to reduce the chance that a given open volume will be efficiently sampled.

Using the crystallographic solution of elastase, in particular, the pig pancreas elastase structural solution of G.A. Petsko of Brandeis University, the method can be exemplified using methanol as the ORF. **Figure 2A** shows the final solution using a relatively high B value, e.g., $B = 10$. **Figure 2B** shows the final solution using an intermediate value, e.g., $B = 6$ or 7 . **Figure 2C** shows the final solution using a lower intermediate value, e.g., $B = 0$ or -2 or -4 . **Figure 2D** shows the final solution using a restrictive value, such as $B = -14$. As illustrated, with lower values of B less and less methanol molecules remain bound. These remaining methanol fragments indicate those that bind with relatively high affinity.

The next step of the process is to conduct simulations with additional ORFs and identify clusters of relatively high affinity ORF binding sites. Thus, for example, again using elastase, simulations can be conducted to determine binding for ORFs for ammonia, methanol, ketone and amide. Combined results at relatively low B values are illustrated in **Figure 3A**. Clusters of ORF binding sites are identified in **Figure 3B**. The method of the present invention seeks to identify clusters of ORF binding sites, where the clusters can be made up solely of one type of ORF. Preferably, however, the cluster will include binding sites for 2, 3, 4, 5, 6, 7 or more distinct ORFs.

Examples of useful ORFs include:

<u>Name</u>	<u>Structure</u>
Acetone	$\text{CH}_3(\text{C}=\text{O})\text{CH}_3$
Aldehyde	$\text{H}(\text{C}=\text{O})-\text{CH}_3$
Amide	$\text{H}(\text{C}=\text{O})\text{NH}_2$
Ammonia	NH_3
Benzene	
Carboxylic Acid	CH_3COOH
1,4-Diazine	

<u>Name</u>	<u>Structure</u>
Ester	$\text{CH}_3\text{-O-(C=O)-CH}_3$
Ether	$\text{CH}_3\text{-O-CH}_3$
Formaldehyde	$\text{H}_2\text{C=O}$
Furan	
Imidazole	
Methane	CH_4
Methanol	CH_3OH
Phospho-Acid	
Pyridine	
Pyrimidine	
Pyrrole	
Thiol	CH_3SH
Thiophene	

Preferably, the ORFs selected are representative of chemical features that have proven useful in the design of pharmaceuticals or other bioactive chemicals.

Thus, in a first mode of analysis, an important part of the process is to run the
 5 simulations with several ORFs, identifying clusters of sites that bind multiple ORFs with

relatively high affinity. These clusters are strong candidate sites for ligand binding sites. Moreover, the relative positioning of the ORFs is instructive of the features of good binding agents. For example, at the binding site identified on elastase by the methods described below, a cluster having two benzene rings with an amide interposed between
5 them models some of the strongest elastase inhibitors derived from an extensive research program, which inhibitors have a sulfonamide in place of the carbon-based amide of the simulation. See, Tables XXIII and XXV of Edwards et al., "Synthetic Inhibitors of Elastase," *Medicinal Research Reviews* 14:127-194, 1994.

In some implementations of the invention, clusters of ORF binding sites alone
10 will identify, or substantially narrow the range of choices for, the sites at which ligands interact with a given protein. However, in some embodiments of the invention, the sites that bind water strongly are identified, and the clusters that intersect with strong water binding sites are discounted. Thus, in the elastase example, the candidate ligand binding sites of **Figure 3B** are narrowed by excluding water binding sites, as illustrated in **Figure**
15 **3C**. If the analysis is extended to five ORFs as illustrated in **Figure 3D**, a single candidate site remains. **Figure 3E** shows a slightly different perspective of the same site illustrated in **Figure 3D**, with the analysis extended to six ORFs. **Figure 3F** shows how well the candidate site (left panel) matches up with the structure of a co-crystal containing the ligand trifluoroacetyl-lysyl-prolyl-*p*-isopropylanilide.

20 Accordingly, in a second mode of analysis, an optional step in the process is to narrow the choices for ligand binding sites by excluding ORF clusters that intersect with relatively strong water binding sites.

It should be noted that clusters of ORFs are typically identified at relatively low *B* values, thereby helping to identify prospective binding sites for ligands. However,
25 further information about prospective binding sites can be gleaned by looking, in the vicinity of a prospective binding site, at more weakly binding ORFs. This information value flows from the prospect of more weakly binding ORFs modeling a ligand interaction which, while weak in isolation, models a real contribution to ligand binding affinity of a bioactive agent as a whole. Illustrated in **Figure 4A** are the amide binding
30 sites extracted from the data of six co-crystallization experiments with elastase and known ligands. Illustrated in **Figure 4B** is a cluster of the highest affinity amide binding sites determined by simulation. Illustrated in **Figure 4C** are the amide ORFs of **Figure**

4B plus amides which are in the vicinity of the cluster but which appear in the simulation at the second highest affinity values. As illustrated, this last step of expanding the results by looking at neighboring lower affinity ORF binding sites helps to better model the results seen in co-crystallography. Specifically, the cluster results identify the site at which the majority of amide binding sites are seen in crystallography, but the expansion extends the results to another cleft in elastase where amides have been experimentally located. Additionally, the expansion identifies part of another cleft at which ligand interactions are seen (as will be illustrated in other Figures).

Thus, in a third mode of analysis, the features of ligand binding sites indicated by other modes of analysis are expanded upon by looking to less stringent simulation results in the vicinity of ORF clusters. The above illustration focused on a cluster of one type of ORF, but is applicable with clusters of many types of ORFs, where the expansions can be limited to one type of ORF or multiple types of ORFs.

The data in **Figures 4A-4C** illustrate an important concept. Both in actual ligand bindings and in the simulations, multiple effective binding locations and orientations for a given type of moiety can be found to overlap. This reflects the existence of multiple local energy minima. In real world actions, rather than low temperature averaging obtained by crystallography, binding interactions will reflect a range of such local minima.

In **Figures 5A** and **5B**, solutions obtained with co-crystals of elastase inhibitors are compared with data obtained by the methods herein described. In **Figure 5A**, the solutions for six co-crystallized inhibitors are shown, with the inhibitor molecules overlaid on each other (non-space-filling representation, with the elastase segment represented by a space-filling illustration). These inhibitors are trifluoroacetyl-l-lysyl-l-prolyl-p-isopropylanilide (crystal solution: Mattos et al., as submitted April 30, 1994), trifluoroacetyl-l-lysyl-l-leucyl-p-isopropylanilide (crystal solution: Mattos et al., as submitted June 22, 1994), trifluoroacetyl-l-phenylalanyl-p-isopropylanilide (crystal solution: Mattos et al., as submitted April 30, 1994), trifluoroacetyl-l-phenylalanyl-l-alanyl-p-trifluoromethylanilide (crystal solution: Mattos et al., as submitted February 14, 1995), trifluoroacetyl-l-valyl-l-alanyl-p-trifluoromethylanilide (crystal solution: Mattos et al., as submitted February 14, 1995) and n-(tert-butoxycarbonyl-alanyl-alanyl)-o-(p-nitrobenzoyl) hydroxylamine (crystal solution: Ding et al., as submitted July 10, 1995).

In **Figure 5B**, the solutions for approximately 10 ORFs, which are in their respective high affinity protein binding states are overlaid. Both methods identify a region which favors the binding of aromatic moieties. The simulation process achieves approximately 90% 3D geometric identity with the crystallography results.

5 **Figures 6A and 6B** show the regions of elastase involved in binding ligands as indicated by the crystallographic data, **Figure 6A**, and as indicated by the solutions obtained from the computational method described herein, **Figure 6B**.

The simulations of the invention utilize a Monte Carlo algorithm. The form of Monte Carlo simulation useful in the present invention is described in Frenkel and Smit,
10 "Understanding Molecular Simulation: From Algorithms to Applications," Academic Press, New York, 1996. The simulation method can comprise:

- Locate a numeric representation of the macromolecule in a periodic cell.
- Optimize the position of the macromolecule in the cell.
- 15 • Locate all the cavities in the macromolecule, whether interior or surface cavities.
- Insert and delete the ORFs (including water) in these cavities.
- Compute the probabilities of occupation of the ORFs using a grand canonical ensemble probability density function.
- 20 • Vary the chemical potential yielding relative free energies of binding.

The methodology, grand-canonical ensemble simulation, can be introduced as follows:

Grand-Canonical Ensemble Simulations

25 The distinguishing feature of simulations in the grand-canonical ensemble is the change in the number of molecules (ORFs) in the system during the simulation. In other words, the sampling is not restricted to the configuration space of a given dimension but it has to be extended to a set of configuration spaces. Applicant has found, unexpectedly, that the complexity of allowing for these changing numbers of molecules and the
30 resulting changing mass nonetheless makes the simulation computationally extremely more efficient. The change in the number of molecules corresponds to the fact that the

grand-canonical partition function Ξ is the linear combination of the corresponding canonical partition functions of a different number, N , of molecules, Q :

$$\Xi(T, V, \mu) = \sum_{N=0}^{\infty} \frac{\exp(\mu N / kT)}{N!} Q(T, V, N) \quad (1)$$

5

where T is the absolute temperature, μ is the chemical potential, k is the Boltzmann constant, and $Q(T, V, N)$ is defined by:

$$Q(T, V, N) = q^N \int \exp(-E(X^N) / kT) dX^N \quad (2)$$

10

with q being the molecular partition function.

The sampling of the configuration space of N molecules (ORFs) has been shown to be feasible using Metropolis Monte Carlo methods where in each step of the sampling a molecule's (ORFs) position is changed by a small amount and the resulting new conformation is accepted or rejected based on the change in energy, ΔE , as a result of the change attempted. This position shifting can be thought of as effecting a "shaking" of the ORF to identify its favored positioning, and the "shaking" methodology, which can be biased in the direction of the forces can be termed "forced bias Monte Carlo." When this shaking is applied, the simulation solutions reflect higher probability orientations.

20 Accordingly:

$$P_{move}^{acc} = \min(1, \exp(-\Delta E / kT)) \quad (3)$$

Notice that the temperature (kept constant during the simulation) enters the acceptance formula as a scaling factor of the energy change.

Generalizing the canonical ensemble Metropolis method to simulations in the grand-canonical ensemble calls for steps where the number of molecules (ORFs) changes. Operationally, this requires either the deletion of an existing molecule or the 'creation', i.e., insertion of a new one. It has been shown that when the deleted molecule is chosen randomly, then the deletion attempt should be accepted with the following probability:

$$P_{del}^{acc} = \min(1, \exp(-\Delta E / kT - B) \frac{N}{V}) \quad (4)$$

where

$$B = \mu' / kT + \ln \langle N \rangle \quad (5)$$

with μ' being the excess chemical potential, N the number of molecules (ORFs), $\langle N \rangle$ its Boltzmann average and V the volume of the system (which is a constant during the simulation). Attempts to insert a molecule (ORF) at a random location is accepted with the following probability:

$$P_{ins}^{acc} = \min(1, \exp(-\Delta E / kT + B) \frac{V}{N+1}) \quad (6)$$

Here the effect of the chemical potential is introduced into the acceptance expression via the B parameter. The presence of the factors V and N follows from the relation between the canonical and grand-canonical partition functions: when a molecule (ORF) is taken out of the system, the integration over its coordinates (in Q_N) will yield a V factor and N is the last factor of $N!$. They can also be given a probabilistic interpretation: the insertion site will be chosen with probability $1/V$ and the molecule (ORF) to be deleted will be chosen with probability $1/N$.

The simulation proceeds by alternating attempts to move, insert and delete molecules (ORFs) and accepting them with probabilities P_{move}^{acc} , P_{ins}^{acc} , P_{del}^{acc} as defined by Equations (3-5) above. After sufficiently long runs, the number of molecules (ORFs) N will fluctuate around its Boltzmann average $\langle N \rangle$. If a given density has to be simulated then it is generally necessary to try different B values. In this regard, it is useful to note the following relationship:

$$\left(\frac{\partial \langle N \rangle}{\partial B} \right)_{T,V} = \langle N^2 \rangle - \langle N \rangle^2 \quad (7)$$

This method has been found useful in simulating atomic fluids at moderate densities but runs into difficulties when room-temperature liquids are simulated. The difficulty stems from the fact that most insertion attempts will be at positions where there already is a molecule (e.g., from the solved protein structure) resulting in a large ΔE , and

5 the resulting probable rejection of the attempt.

To increase the efficiency of insertion attempts, a cavity-biased insertion technique was introduced. Insertions are attempted only at sites where a cavity of suitable size already exists, thereby ensuring a non-negligible probability of acceptance. However, to ensure that the simulation thus modified still produces the required

10 Boltzmann distribution, both the insertion and deletion acceptance probabilities have to be modified. The modified expression involves the probability of finding a cavity when there are N molecules (ORFs) present, P_N^{cav} , which follows from:

$$P_{ins}^{acc} = \min(1, \exp(-\Delta E / kT + B) \frac{P_N^{cav} V}{N + 1}) \quad (8)$$

15

$$P_{del}^{acc} = \min(1, \exp(-\Delta E / kT - B) \frac{N}{P_N^{cav} V}) \quad (9)$$

In Equations 8 and 9, $P_N^{cav} N$ represents the volume of the regions of the system that contain cavities of suitable size. The efficiency of the cavity-biased method follows from

20 the fact that the algorithm searching for cavities also yields P_N^{cav} without extra steps. Calculations on a variety of fluids (water, benzene), which define ORFs, have confirmed that the cavity biased method significantly increases the efficiency of insertion attempts and allows modeling of densities that proved to be impractical without this improvement.

Water binding

Aspects of the simulations used in the invention can be illustrated with calculations used to determine the strength of water binding to a synthetic polynucleotide.¹ This illustration can be described as follows:

- 5 This text illustrates how the method of *simulated annealing of chemical potential* allows bulk waters to be distinguished from bound waters, and how differentially bound waters may be distinguished from each other based on their relative chemical potentials. This is illustrated by showing that it takes more free energy to desolvate the minor groove than the major groove of a charged DNA dodecamer.
- 10 Grand canonical ensemble simulations are generally performed by placing a molecule in a periodic simulation cell, setting a parameter B , which is representative of free energy, in such a way as to achieve an experimentally determined density, sampling potential hydration positions around the molecule by inserting and deleting water molecules from the simulation cell using a technique such as cavity-bias,^{2,3} and accepting
- 15 or rejecting the attempt based on a Metropolis Monte Carlo⁴ criteria using a grand canonical ensemble probability function.⁵ The parameter B is related to the excess chemical potential μ' as follows: $B = \mu'/kT + \ln\langle N \rangle$, where k is Boltzmann's constant, T is the absolute temperature, and $\langle N \rangle$ is the mean number of molecules of the ORF, which here is H₂O. In the method of *simulated annealing of chemical potential*, the
- 20 simulation is started with a large initial B -value so that a higher percentage of water insertion attempts are accepted. This causes the simulation cell to be flooded with water molecules. After this grand canonical ensemble simulation at high excess chemical potential is equilibrated, subsequent simulations are carried out at successively lower B -values. This successive lowering of the B -values causes a gradual removal of the bulk
- 25 water molecules from the simulation cell. As the chemical potential is further "annealed", a point is reached at which water molecules do not readily leave the cell, thereby identifying those water molecules that are strongly influenced by the DNA, the so-called "bound water molecules". As the excess chemical potential is again lowered, ultimately some of these bound waters start to leave the cell. Since chemical potential is
- 30 a free energy, this *simulated annealing of chemical potential* yields a numerical estimate of the differential free energy of binding of the different bound water molecules. It must be emphasized that our utilization of the term "annealing" applies strictly to the value of

the chemical potential and that the temperature is kept constant at, for example, 298 K in all the simulations. For all simulations the DNA was held fixed, water molecules were added and deleted throughout all parts of the cell, extensive canonical Monte Carlo was performed between accepted grand canonical Monte Carlo steps, and periodic boundary conditions were used.

As an illustration of the method, a *simulated annealing of chemical potential* on a d(CGCGAATTCGCG)₂ was performed, starting with $B = 1.0$ down to -26 in 37 increments performing 2,000,000 cavity-biased grand canonical ensemble Monte Carlo steps at each B -value. The final configuration of the simulation with $B = -6$, has 1120 bound water molecules. The final configuration of the simulation with $B = -8$, has bound 533 water molecules. The final configuration of the simulation with $B = -9$, has 390 bound water molecules. The final configuration of the simulation with $B = -11$, has 215 bound water molecules. The most salient feature of this progression is the differential hydration of the major and minor groove of the DNA. The $B = -6$ simulation shows the DNA essentially uniformly solvated. The $B = -8$ simulation clearly shows that upon lowering of the chemical potential by 2 B-units, a majority of the nonbulk extracted waters come from the major groove, while the minor groove remains almost unaffected. Annealing the chemical potential further ($B = -9$) still leaves the minor groove well hydrated while the major groove is almost stripped. Lowering B even further ($B = -11$) results in the removal of almost all water molecules from both the major and minor groove. Quantitation of the hydration of the DNA as a function of chemical potential was computed by proximity analysis^{6,7} with the results shown in Table 1:

B	<u>first hydration shell</u>				<u>first and second hydration shell</u>			
	<u>minor groove</u>		<u>major groove</u>		<u>minor groove</u>		<u>major groove</u>	
	no. of waters	density	no. of waters	density	no. of waters	density	no. of waters	density
-6	7.27	0.013	13.23	0.012	21.3	0.021	41.7	0.011
-8	5.4	0.010	5.06	0.004	14.6	0.015	11.8	0.003
-9	4.08	0.007	4.36	0.004	11.5	0.011	9.7	0.003
-11	1.04	0.002	2.11	0.002	3.9	0.004	4.2	0.001

For $B = -6$, the first hydration shell (defined by the position of the first minimum of the radial distribution function) of the major and minor groove has a comparable density (0.012 and 0.013, respectively), while the second hydration shell of the minor groove has

twice the density of the major groove. For $B = -8$ the hydration difference becomes quite pronounced with the minor groove first and second shell hydration density being 2.5 fold and 5 fold higher than the major groove, respectively. For $B = -11$ the major and minor groove hydration density again becomes equal because at this value of the
5 excess chemical potential both grooves are essentially stripped bare.

Illustrating the differential hydration propensities of the major and minor grooves of DNA is computationally undemanding (3 days of CPU time to run one annealing schedule and 3 days of CPU time to run one proximity analysis⁷ on an SGI Power Challenge) using *simulated annealing of chemical potential* because only a coarse
10 "cooling" schedule of the chemical potential is required. Since the chemical potential is a free energy, a very fine cooling schedule may be used to estimate quantitatively the hydration free energy difference of two different functional groups or even two different atoms of the DNA. Two atoms that desolvate at the same B -value have similar solvation free energy, or alternatively, require a finer cooling schedule to resolve the differences. It
15 should be noted that the model system used here consisted of ionic DNA with 22 negative charges and no sodium counterions. The findings presented herein about the preferential hydration of the minor groove corresponds very well to results from X-ray crystallographic and NMR studies. Possible reasons for the stronger binding of water molecules in the minor groove may include the following: the high density of the
20 charged rows of phosphate groups, steric constraints, and specific water—water, water—DNA interactions.

The regions where water binds tightly on a protein, are regions which are precluded from ORF binding. Thus, the remaining sites on the protein unoccupied by water are candidates for good ORF binding.

25

Antagonists and Agonists - Assays and Molecules

Candidate bioactive agents identified by the methods of the invention can be tested to assess their binding to the macromolecule in question. Where the macromolecules are responsible for many biological functions, including disease states, it is therefore desirable to
30 devise screening methods to identify compounds which stimulate or which inhibit the function of the macromolecule. Accordingly, in a further aspect, the present invention provides for a method of screening compounds to identify those which stimulate or which

inhibit the function of such a macromolecule. In general, agonists or antagonists can be employed for therapeutic and prophylactic purposes for diseases. Compounds can be identified from a variety of sources, for example, cells, cell-free preparations, chemical libraries, and natural product mixtures.

- 5 The screening methods can simply measure the binding of a candidate compound to the macromolecule, or to cells or membranes bearing the macromolecule. The macromolecule can be a variant of the macromolecule used in the simulation method, such as a fragment retaining the binding site identified in the simulation or a fusion protein used to make recombinant synthetic methods more practical. The screening method can involve
- 10 competition with a labeled competitor. Further, these screening methods can test whether the candidate compound results in a signal generated by activation or inhibition of the macromolecule, using detection systems appropriate to the cells comprising the macromolecule. Inhibitors of activation are generally assayed in the presence of a known agonist and the effect on activation by the agonist by the presence of the candidate
- 15 compound is observed. Further, the screening methods can simply comprise the steps of mixing a candidate compound with a solution containing a macromolecule, measuring macromolecule activity in the mixture, and comparing the activity of the mixture to a standard.

- The invention also provides a method of screening compounds to identify those
- 20 which enhance (agonist) or block (antagonist) the action of macromolecules, including association of the macromolecule with itself or another macromolecule. The method of screening can involve high-throughput techniques. For example, to screen for agonists or antagonists, a synthetic reaction mix, a cellular compartment, such as a membrane, cell envelope or cell wall, or a preparation of any thereof, comprising macromolecule and a
- 25 labeled substrate or ligand of such polypeptide is incubated in the absence or the presence of a candidate molecule that can be a agonist or antagonist. The ability of the candidate molecule to agonize or antagonize the macromolecule is reflected in decreased binding of the labeled ligand or decreased production of product from a substrate. Molecules that bind gratuitously, *i.e.*, without inducing the effects of macromolecule are most likely to be good
- 30 antagonists. Molecules that bind well and, as the case can be, increase for example the rate of product production from substrate, increase signal transduction, or increase chemical channel activity are agonists. Detection of the rate or level of, as the case can be, production of

product from substrate, signal transduction, or chemical channel activity can be enhanced by using a reporter system. Reporter systems that can be useful in this regard include but are not limited to colorimetric, labeled substrate converted into product, a reporter gene that is responsive to changes in macromolecule activity, and binding assays known in the art.

5

References

All publications and references, including but not limited to patents and patent applications, cited in this specification are herein incorporated by reference in their entirety as if each individual publication or reference were specifically and individually indicated to be incorporated by reference herein as being fully set forth. Any patent application to which this application claims priority is also incorporated by reference herein in its entirety in the manner described above for publications and references. The cited documents incorporated by reference into this disclosure include:

- ¹ Guarnieri and Mezei, *J. Am. Chem. Soc.* 118: 8493-8494, 1996.
- ² Mezei, *Mol. Phys.* 61: 565-582, 1994.
- ³ Resat and Mezei, *J. Am. Chem. Soc.* 116: 7451-7452, 1994.
- ⁴ Metropolis et al., *J. Chem. Phys.* 21: 1087-1092, 1953.
- ⁵ Tolman, R. In *The Principles of Statistical Mechanics*, Dover Press, New York, 1971.
- ⁶ Mehrotra and Beveridge, *J. Am. Chem. Soc.* 102: 4287, 1980.
- ⁷ The effects of different partial charges on proximity analysis are described in: Mezei, *Mol. Simul.* 1: 327-332, 1988.
- ⁸ Calculations of volume elements can be CPU intensive. The effects of volume element calculations on proximity analysis are described in Mezei and Beveridge In *Methods in Enzymology*, Packer, Ed., Academic Press, New York, 1986, pp. 21-47.

25

While this invention has been described with an emphasis upon preferred embodiments, it will be obvious to those of ordinary skill in the art that variations in the preferred devices and methods may be used and that it is intended that the invention may be practiced otherwise than as specifically described herein. Accordingly, this invention includes all modifications encompassed within the spirit and scope of the invention as defined by the claims that follow.

30